FastDiff: A Fast Conditional Diffusion Model for High-Quality Speech Synthesis

Year: 2022 | Citations: 204 | Authors: Rongjie Huang, Max W. Y. Lam, J. Wang, Dan Su, Dong Yu

Abstract

Denoising diffusion probabilistic models (DDPMs) have recently achieved leading performances in many generative tasks. However, the inherited iterative sampling process costs hindered their applications to speech synthesis. This paper proposes FastDiff, a fast conditional diffusion model for high-quality speech synthesis. FastDiff employs a stack of time-aware location-variable convolutions of diverse receptive field patterns to efficiently model long-term time dependencies with adaptive conditions. A noise schedule predictor is also adopted to reduce the sampling steps without sacrificing the generation quality. Based on FastDiff, we design an end-to-end text-to-speech synthesizer, FastDiff-TTS, which generates high-fidelity speech waveforms without any intermediate feature (e.g., Mel-spectrogram). Our evaluation of FastDiff demonstrates the state-of-the-art results with higher-quality (MOS 4.28) speech samples. Also, FastDiff enables a sampling speed of 58x faster than real-time on a V100 GPU, making diffusion models practically applicable to speech synthesis deployment for the first time. We further show that FastDiff generalized well to the mel-spectrogram inversion of unseen speakers, and FastDiff-TTS outperformed other competing methods in end-to-end text-to-speech synthesis. Audio samples are available at https://FastDiff.github.io/.